빅콘테스트 2016 챌린지 리그

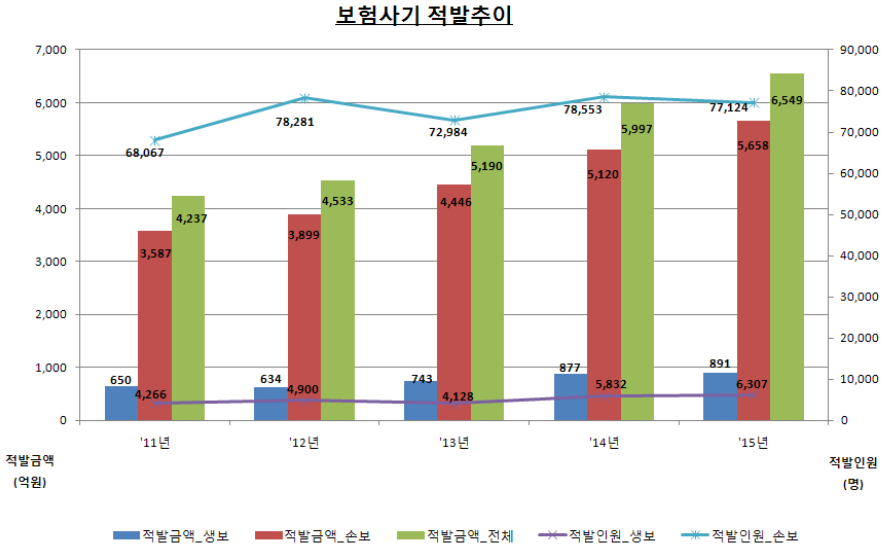
- 빅데이터 분석을 통한 보험사기 예측 알고리즘 개발 -

고려대학교 산업경영공학 대학원 서덕성
연세대학교 정보산업공학 대학원 이경택
인하대학교 통계학과 김상진
인하대학교 통계 대학원 박희경

INDEX

01 Data Explore

보험사기 현황 및 문제점



※ 출처 : 금융감독원 보험사기방지센터

매년 보험사기 적발 인원(2015년에서 전년 대비 1.1% 감소)은 상승세가 별로 없는 반면, 적발금액은 매년 높은 상승세(2015년에서 전년 대비 9.2% 증가)를 보인다. 이에 대해, 보험사기의 적발은 잘 안 되지만, 보험사기의 정도가 점점 심해지고 있다고 생각할 수 있다.

하지만, 현재 보험사기를 적발하는 데에는 여러 가지 손해 및 한계(인력 부족, 비용의 한계, 불확실성 등)에 의해 어려움에 봉착해있다.

01 Data Explore

주제 정의 및 데이터 소개

: 매년 증가하는 보험사기(보험사고의 발생, 원인 또는 내용에 관하여 보험자를 기망하여 보험금을 청구하는 행위)를 고객 및 보험 관련 데이터를 통해, 효율적으로 예측하는 알고리즘 개발

사용 Data

- CUST (고객의 특성 Data)
- CNTT (고객들의 계약 속성 Data)
- CLAIM (고객들을 대상으로 한 지급 속성 Data)
- FMLY (고객간 가족여부 Data)
- FPINFO (보험설계사 Data)

분석을 하기 위한,
데이터 정제 및 JOIN 필요

01 Data Explore

파생변수 생성 - ① 범주형 변수

데이터에서 보험이 여러 개가 있는 ID가 있으므로, 범주형의 변수 처리를 COUNT 형식으로 파생변수를 생성하였다. 또한, 변수에 NA가 있으면 새로운 범주형을 만들어 코딩하였다.

Ex) 한 고객(CUST_ID=10)이, 재해에 대한 사고(ACCI_DVSN)를 3번 당했을 때, 파생변수 사고구분_재해에 대해서 “3”의 값을, 사고구분_교통재해, 질병, NA에 대한 값은 “0”으로 처리
한 고객 (CUST_ID=13)이, 질병이 2번, 교통재해를 1번 당했을 때, 파생변수_교통재해의 값을 “1”, 파생변수_질병의 값을 “2”로 처리한다.

CUST_ID	사고구분 (ACCI_DVSN)
10	1 (재해)
10	1 (재해)
10	1 (재해)
11	2 (교통재해)
12	1 (재해)
13	3 (질병)
13	3 (질병)
13	2 (교통재해)

CUST_ID	사고구분_재해	사고구분 _교통재해	사고구분_질병	사고구분_NA
10	3	0	0	0
11	0	1	0	0
12	1	0	0	0
13	0	1	2	0

01 Data Explore

파생변수 생성 - ② 수치형 변수

데이터에서 보험이 여러 개가 있는 ID가 있으므로, 수치형 변수 처리를 그 ID를 대표할 수 있는 MAX, MIN, MEAN으로 파생변수 생성.
단, 한 ID에 여러 개의 데이터가 있으면, 부분에 NA가 존재한다면 무시하고 계산.
한 ID에 한 개의 데이터가 있는데, NA라면 “-100”으로 처리.

Ex) 고객의 지급금액에 대한 파생변수로, 고객의 지급금액_MAX, MIN, MEAN으로 파생변수 생성

CUST_ID	지급금액 (PAYM_AMNT)
10	600,000
10	31,900,000
10	270,000
11	40,000
12	72,320
13	170,000
13	425,000
13	310,000

CUST_ID	지급금액_MAX	지급금액_MIN	지급금액_MEAN
10	31,900,000	270,000	10,923,333.33
11	40,000	40,000	40,000
12	72,320	72,320	72,320
13	425,000	170,000	301,666.67

01 Data Explore

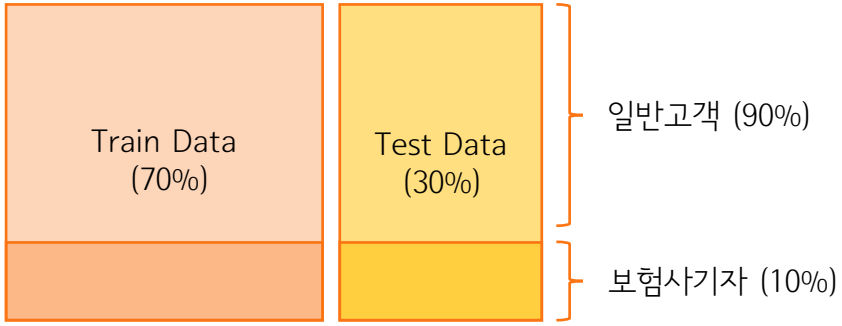
데이터 조인

CNTT(계약 Data), CLAIM(지급 속성), FMLY(가족 속성) Data를 고객 ID(CUST_ID)의 기준으로 CUST(고객 Data)와 함께 JOIN 하여 분석용 데이터 생성.

CUST_ ID	기존 변수 + 파생 변수
1 2 ⋮ 20,606 20,607	**20,607 X 4,017 Data Frame**

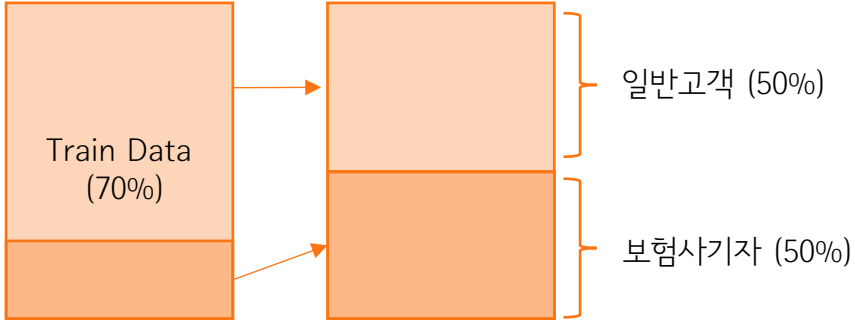
02 Modeling

Model1 – 단일 모델 부스팅



전체 데이터에서 일반고객과 보험사기자의
비율(약 9:1)을 유지하며 Train : Test 데이터를
7 : 3으로 나누어, 앙상블(Ensemble)기법 중 하나인
부스팅(Boosting)기법을 통하여 예측
→ F-measure : 0.5873

Model2 – Up&Down Sampling을 이용한 부스팅의 배깅

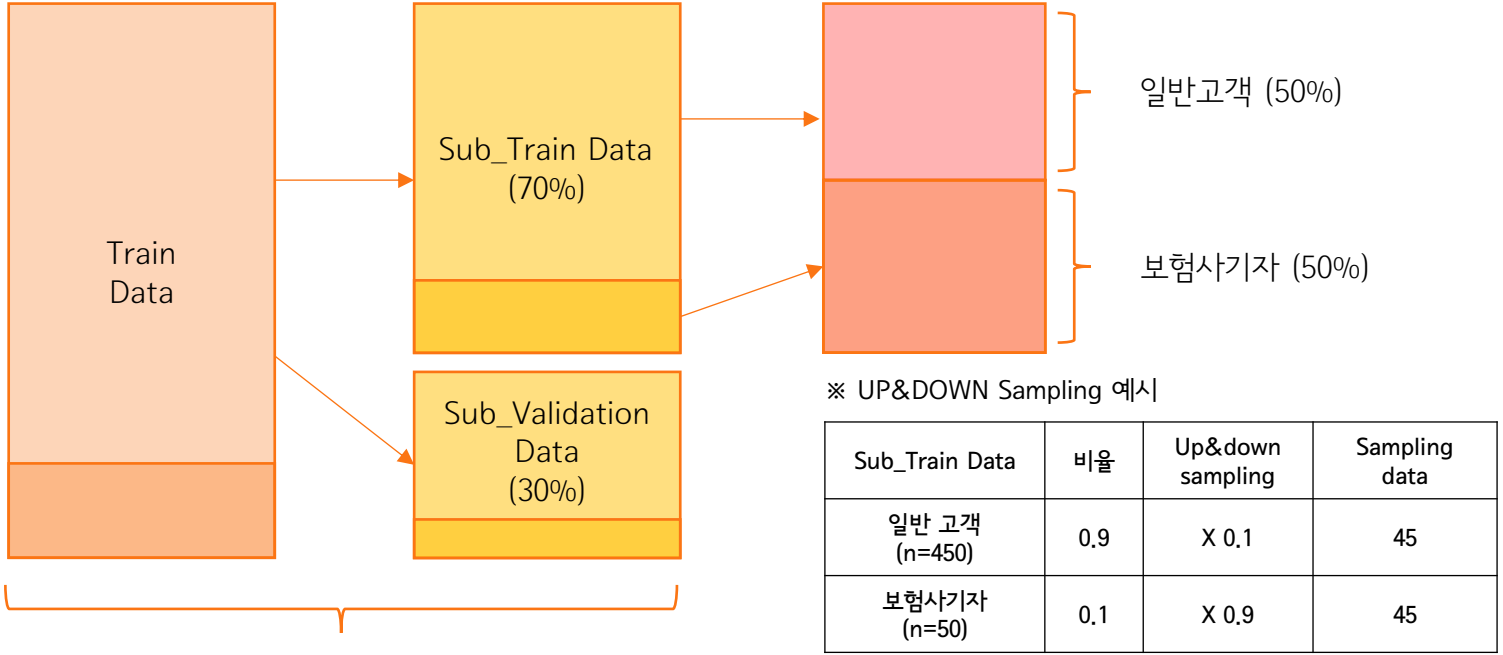


전체 중 오직 10%만이 보험사기자인 불균형문제를
해결하기 위해, 일반고객과 보험사기자의 비율을
5:5로 맞추는, Up&down Sampling을 시행하여,
부스팅을 여러 번 시행하여 예측. (일반고객에 대해서는
Down, 보험사기자에 대해서는 Up Sampling을 적용)
→ F-measure : 0.6211

02 Modeling

Model3 – F-measure weighting을 이용한 Gradient Boosting Simulation

Step1> Train Data를 일반고객과 보험사기자 비율을 유지하여 다시 7:3으로 나눈다.
(이를 Sub_Train Data, Sub_Validation Data라고 정의한다.) 이후, Sub_Train Data에서 일반고객과 보험사기자의 비율을 5:5로 맞추기 위해 Up&Down Sampling을 실시한다.



모든 Data에서 일반고객과 보험사기자의 비율은 9:1

※ UP&DOWN Sampling 예시

Sub_Train Data	비율	Up&down sampling	Sampling data
일반 고객 (n=450)	0.9	X 0.1	45
보험사기자 (n=50)	0.1	X 0.9	45

관측자료의 개수가 줄어드는 DOWN Sampling 문제 해결

02 Modeling

Model3 – F-measure weighting을 이용한 Gradient Boosting Simulation

Step2> 일반고객과 보험사기자의 비율을 1:1로 맞춘 후(Step1), 변수를 ½개 무작위 선택하고, 이후 Gradient boosting으로 모델을 구축하여, Sub_Validation Data에 예측 적합하여 F-measure를 기록하여 선택된 변수에 할당하고, 선출되지 않은 변수들은 0의 값을 할당한다.

Step3> Step2를 총 500회 반복하여 각각의 F-measure를 기록한 후 500번의 Simulation 기록의 평균으로 변수의 가중치를 부여한다.

※ Gradient Boosting Simulation 예시

Simulation	변수 1	변수 2	변수 3	…	변수 n
F-measure1	0.593	0.593	0	…	0
F-measure2	0.573	0	0.573		0.573
⋮	⋮				⋮
F-measure500	0	0.456	0.456	…	0
Mean(F-measure)	0.571	0.552	0.521	…	0.53

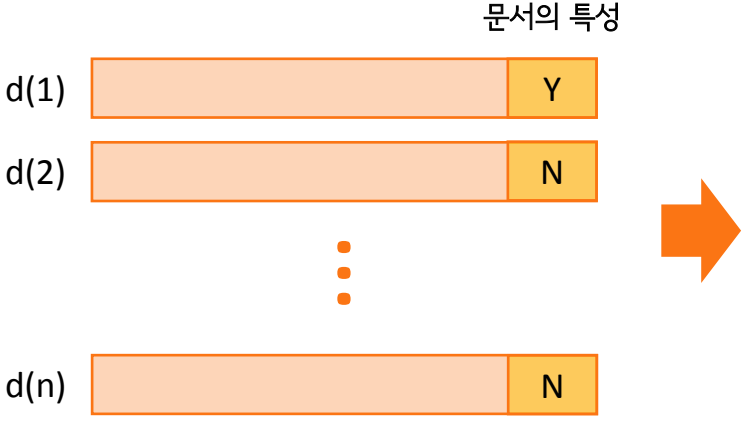
Step4> Simulation을 통해 얻어진 변수의 가중치(Step3)를 통해, 부스팅을 이용하여 예측 → F-measure : 0.6341

02 Modeling

Model4 – Text mining의 prob-weighting을 이용한 Gradient Boosting Simulation

※ Text mining의 prob-weighting 이란?

Step1> 각 문서(Document)를 행으로, 단어(Term)를 열로 가지는 행렬로, 해당 문서에 해당 단어가 포함된 유무 (문서에 단어가 존재 – "1" / 존재하지 않으면 –"0") 를 나타내는 Document-Term Matrix를 만든다.
이때, 각 문서에는 문서의 해당하는 이진분류의 특성값이 있다.(ex. 긍정/부정, Yes/No)



	Term 1	Term 2	…	Term v
Document 1	1	1	…	0
Document 2	1	0		1
⋮	⋮			⋮
Document n	0	1	…	0

(n X v) size – sparse data

02 Modeling

Model4 – Text mining의 prob-weighting을 이용한 Gradient Boosting Simulation

※ Text mining의 prob-weighting 이란?

Step2> 각 단어에 대해서, 문서의 특성과 2X2 분할표를 만들어, prob-weight를 계산한다.

(단, $prob\ weight_i = \log\left(1 + \frac{a_i}{b_i} \cdot \frac{a_i}{c_i}\right)$)

Term 1	문서 Y	문서 N
단어 1	a_1	b_1
단어 0	c_1	d_1

prob-weight$_1$

• • •

Term v	문서 Y	문서 N
단어 1	a_v	b_v
단어 0	c_v	d_v

prob-weight$_v$

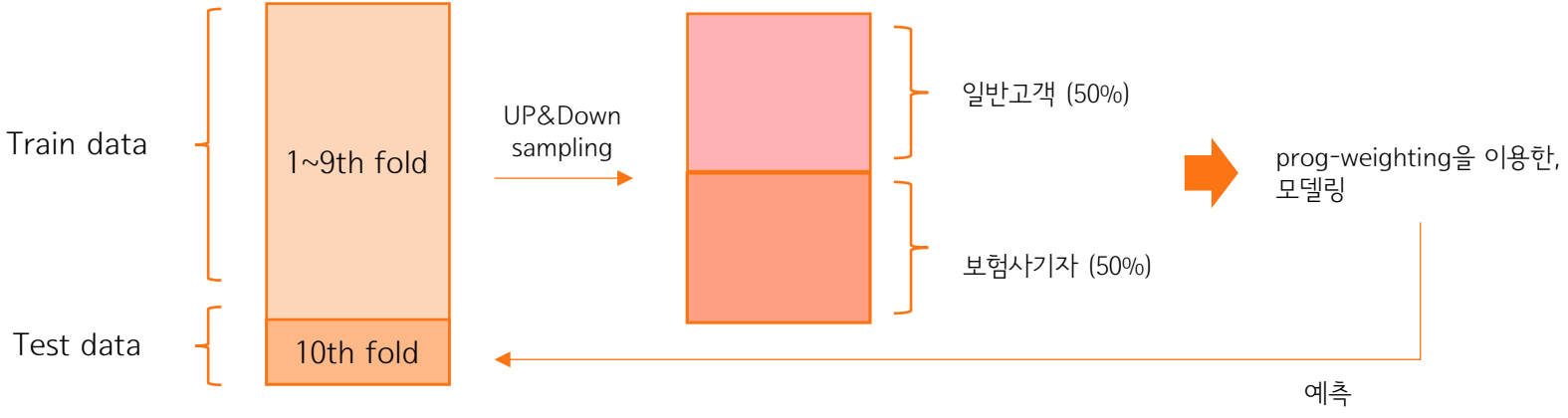
02 Modeling

Model4 – Text mining의 prob-weighting을 이용한 Gradient Boosting Simulation

Step1> Text mining에서 "문서→고객, 단어→변수, 문서의 특성→사기 유무"로 착안한다. 이때, 변수 중 이진분류가 아닌 연속형인 경우 변수의 "평균 이상 : 1 / 평균 이하 : 0"으로 이진 변수로 변형해준다.

Step2> 10 Cross-Validation의 Train data에서 각 변수의 prob-weighting를 구하여, 변수의 가중치로 부여한다.

Step3> Train data를 UP&DOWN sampling과 prob-weighting을 통해 데이터와 변수를 선택해, 20번 modeling을 하여 Test data를 예측한다. (Step1~Step3를 10번 반복) → F-measure : 0.6510



02 Modeling

Model5 – 데이터 Scoring을 통한 모델 안정성 확보

※ 기존 Cross Validation에서의 문제점 (ex. 5-fold Cross Validation_1번째 Simulation)

CV	Fold 1	Fold 2	Fold 3	Fold 4	Fold5
F-measure	0.58	0.68	0.69	0.71	0.64

F-measure가 유독 낮은 Fold가 존재 및 낮은 F-measure의 편차가 심하다.

→ 이유 : F-measure가 낮은 Fold의 나머지 Fold(Train data)의 샘플이 좋지 않다.
(ex. Outlier 및 noise가 많이 있다.)

해결방법

Step1> 각 Fold에 있는 관측치에, Fold에 해당하는 F-measure를 할당한다. (10번 Simulation 시행)

Step2> 각 관측치의 가중치로 관측치의 해당 F-measure의 평균을 이용한 가중치를 할당하여, 복원 5-fold Cross Validation을 통해 모델링을 한다.

02 Modeling

Model5 – 데이터 Scoring을 통한 모델 안정성 확보

※ Sampling Example

고객	Simulation 1	Simulation 2	...	Simulation 9	Simulation 10	F-measure 평균
Obs 1	F-measure 1_1(0.52)	F-measure 2_2		F-measure9_3	F-measure10_1	F1
Obs 2	F-measure 1_2(0.62)	F-measure 2_1	…	F-measure9_1	F-measure10_1	F2
Obs 3	F-measure 1_1(0.52)	F-measure 2_5		F-measure9_3	F-measure10_2	F3
Obs 4	F-measure 1_4(0.60)	F-measure 2_2		F-measure9_2	F-measure10_5	F4
⋮						
Obs n	F-measure 1_1(0.52)	F-measure 2_1	…	F-measure4_4	F-measure10_3	Fn

※ F-measure a_b : a번째 simulation의 해당 고객이 있는 b-fold의 F-measure (ex. F-measure2_1 : 2번째 simulation의 1-fold의 F-measure)

각 관측치의 가중치

$weight_i = 1 - \frac{f_i - \min(f_i)}{\max(f_i) - \min(f_i)}$ 를 sigmoid 함수를 통해서, weight의 분산을 크게 했다.

(각 변수의 weight를 넓게 퍼트려, Sampling 할 때 모형 적합에 좋은 관측치는 더 잘 뽑히고, 나쁜 관측치는 적게 뽑아, Sampling의 효과를 극대화하기 위해서 → 모형의 안정성 확보)

02 Modeling

Model5 – 데이터 Scoring을 통한 모델 안정성 확보

※ 가중치 Sampling의 효과

가중치 Sampling 사용 이전

CV	Fold 1	Fold 2	Fold 3	Fold 4	Fold5
F-measure	0.60	0.63	0.69	0.73	0.69

평균 : 0.668

문제점 : 비정상적으로 낮은 F-measure가 존재하고, F-measure의 편차가 심하다.

가중치 Sampling 사용 이후

CV	Fold 1	Fold 2	Fold 3	Fold 4	Fold5
F-measure	0.68	0.66	0.69	0.70	0.67

평균 : 0.68

효과 : 전체적으로 F-measure가 상승하며, F-measure의 편차가 완화되어 안정성을 확보하였다.

→ F-measure : 0.6819

03 Model Assessment

F-measure를 통한 모델 비교

	F-measure
Model1	0.5873
Model2	0.6211
Model3	0.6341
Model4	0.6510
Model5	0.6819

- Up&Down Sampling을 사용하지 않은 Model1의 효과가 유독 좋지 않은 것을 보아, 일반 고객과 보험사기자의 비율을 맞추어준 UP&DOWN Sampling의 성능이 좋다고 생각된다.

- 변수의 가중치를 조절하지 않은 (Model1, Model2) 과 변수의 가중치를 조절한 (Model3, Model4, Model5)를 전체적으로 비교했을 때 변수의 가중치를 통한 부스팅의 효과가 좋았다.

- Cross-Validation에서 관측치들의 가중치를 이용해서 Sampling한 Model5가 가장 높은 F-measure를 보인다.

F-measure가 가장 높은 Model 5 채택

04 Summary

① Up&Down Sampling을 통한 Class imbalanced data의 문제점 보완

② Ensemble 기법을 통한, 모형의 예측력 및 정확도 향상

③ prob-weighting 및 Boosting을 활용한 변수 선택

④ Data 별 Scoring(Data 별 Sampling probability를 부여)을 통한 모형 안정성 확보

최종 모형의 10 Cross-validation의 F-measure : 0.6819

Thank You